Soc. Netw. Anal. Min. (2016) 6:68
DOI 10.1007/s13278-016-0385-2

ORIGINAL ARTICLE

# Toward understanding online sentiment expression: an interdisciplinary approach with subgroup comparison and visualization

**Bo Gao[1] · Bettina Berendt[1] · Joaquin Vanschoren[2]**

Received: 18 December 2015 / Revised: 23 May 2016 / Accepted: 25 August 2016


**Abstract** Understanding users' sentiment expression in social media is important in many domains, such as marketing and online applications. Is one demographic group inherently different from another? Does a group express the same sentiment both in private and public? How can we compare the sentiments of different groups composed of multiple attributes? In this paper, we take an interdisciplinary approach toward mining the patterns of textual sentiments and metadata. First, we look into several existing hypotheses in social science on the interplay between user characteristics and sentiments, as well as the related evidence in the field of social network data analysis. Second, we present a dataset with unique features (Facebook users chats and posts in multiple languages) and a procedure to process the data. Third, we test our hypotheses on this dataset and interpret the results. Fourth, under the subgroup discovery paradigm, we present an approach with two algorithms that generalizes single-attribute testing. This approach provides more detailed insight into the relationships among attributes and reveals interesting attribute value combinations with distinct sentiments. It also offers novel hypotheses for examination in future studies. Fifth, because the number of mined subgroup comparisons can be large, we develop an exploratory visualization tool that summarizes the comparisons and highlights meta-patterns.

**Keywords** Online social network · Sentiment · Subgroup comparison · Information visualization · Exploratory data analysis

## 1 Introduction

Understanding users' sentiments in social media is important in many domains, such as marketing, sociological/psychological study and online application development. For example, in marketing, data analysts monitor and mine texts in social media to discover how participants in specific demographic groups react to certain brands or events. An analyst must be aware of existing sentiment differences. For instance, do older people express themselves more positively? Is there a difference in sentiment expression between married and single people? However, most hypotheses are based on offline studies. It is thus interesting to test and examine them in more detail with online social network data.

Recently, there has been a large interest in Facebook sentiment analysis Kramer et al. (2014), Siganos et al. (2014). To the best of our knowledge, all the existing sentiment analysis has been conducted on status updates, or other (semi-)publicly available data in online social networks. Users utilize different privacy settings to post or chat on social networks. Is there a sentiment difference between different privacy settings? In this paper, we discover and compare the sentiment patterns in both posts and chats on Facebook in a more differentiated way. Furthermore, most studies have focused on the correlations

✉ Bo Gao
bo.gao111@gmail.com

Bettina Berendt
bettina.berendt@cs.kuleuven.be

Joaquin Vanschoren
j.vanschoren@tue.nl

[1] Department of Computer Science, KU Leuven, Leuven, Belgium

[2] Department of Mathematics and Computer Science, TU Eindhoven, Eindhoven, Netherlands

Published online: 02 September 2016

between single (demographic) factors and sentiments in online social networks. It is potentially more productive to study the sentiment differences using multiple factors. For example, the male users of 21–24 years old with the "friends" privacy setting (see Sect. 3) are less positive than those of 25–28 year old. This type of pattern mining falls under the subgroup discovery paradigm. We propose two algorithms to extract subgroup comparisons of differentiated sentiments.

Although the extraction of subgroup comparisons helps us gain insight into the patterns of OSN users' sentiment expression, the number of extracted patterns could be large, preventing us from understanding the patterns on the meta-level. To address this issue, we develop an exploratory visualization tool that summarizes the comparisons and highlights meta-patterns.

The remainder of the paper is organized as follows: In Sect. 2, we explicate our research questions in light of relevant literature. In Sect. 3, we describe our dataset. In Sect. 4, we test the sentiment differences for single-attribute subgroups. In Sect. 5, we detail our approach to discover "interesting" multi-attribute subgroups. In Sect. 6, we detail the development of the visualization tool, motivate our design choices and demonstrate the potential usefulness of the tool. We summarize the paper in Sect. 7. In Sect. 8, we discuss the limitations and the outlook of our approach.

## 2 Related work and research questions

We use the term "sentiment" to refer to a simplified attitude or emotional state that can be characterized as positive, negative or neutral. For a given document, the positive sentiment strength is $s^+ > 0$, its negative sentiment strength is $s^- < 0$, and we consider its expressiveness to be $(s^+ - s^-)$. We note that the gender-wise online sentiment differences have been extensively studied (e.g., Thelwall et al. 2010b); thus, we will not investigate this in single-attribute hypothesis testing, but we will see the interactions of "gender" with other attributes in Sect. 5.

When a user posts or chats on Facebook, each post or chat has an audience range, mostly definable by the user. For example, a post's privacy setting can be adjusted from only visible to oneself to the entire Web. The number of participants also implicitly defines the audience range of a private chat. Often, people express themselves positively rather than negatively on Facebook, as negative emotions are not socially favorable and people tend to suppress negative emotions in public Gross et al. (2006). Different levels of privacy settings may trigger different sentiment expressions. Our corresponding research questions are:

> RQ1: In Facebook chats and posts, (a) do users express more positive and/or less negative sentiment in public than in private? (b) Is there a difference in expressiveness?

For age-differentiated emotional behavior, Gross et al. (1997) investigated subjects' emotional experience, expression and control. The results consistently showed that, compared to younger subjects, older subjects reported fewer negative emotional experiences and greater emotional control. Furthermore, Stone et al. (2010) found that people's positive emotional state increases after 50 years old. Stress and anger steeply decline from the early 20 s. Worry was elevated through middle age (30–59 years old) and then declined. However, do these findings translate to the communication in online social networks? Our corresponding research questions are:

> RQ2: In Facebook chats and posts, (a) Does negative sentiment decline after people's early 20 s but increase during middle age? (b) Do people older than 50 years old express themselves more positively?

Researchers have also studied emotional differences in terms of relationship status. For instance, Yap et al. (2012) found that married people reported higher levels of satisfaction than they did while being single. Another study Taylor (2009) showed that being in a relationship was associated with higher levels of anger. Our corresponding research questions are:

> RQ3: In Facebook chats and posts, (a) do married people express more positive and/or less negative sentiment than single people? (b) Do people in a relationship (not married) express less positive and/or more negative sentiment than single people? (c) Is there a difference between the people who are married and those that are in a relationship?

Moreover, the comparison of sentiment differences between single-attribute subgroups can be generalized to multi-attribute subgroups. This falls in field of subgroup discovery (SD). The concept of SD was initially introduced by Klösgen (1996) and Wrobel (1997). SD is a set of techniques that are useful in exploratory data analysis. Unlike conventional supervised or unsupervised learning that often builds a global model with an optimization criterion, SD lays emphasis on describing partial relationships between data attribute values of interest and target variable(s). Also different from such transformative data preprocessing techniques as feature selection and dimensionality reduction, SD requires the original attributes in a dataset to be retained, as it is the relationships between these attributes and the target variable(s) that are of interest.

The data to be analyzed by SD are in tabular form. The task of SD is to discover subgroups of a given dataset that are statistically "interesting." In other words, subgroups need to have "unusual" distributional characteristics with respect to the corresponding attribute(s) of interest.

Recent advances in the field of subgroup discovery enable fast discovery of "quality" subgroups with high diversity and low redundancy van Leeuwen and Knobbe (2012), or discover subgroups with multiple numerical target attributes Leman et al. (2008). To the best of our knowledge, existing approaches extract subgroups that have unusual or distinct distributional characteristics with respect to the entire population. However, instead of individual subgroups, we are interested in "local comparisons" between subgroups, with one or more attributes. Our corresponding research question is:

> RQ4: How can we discover "interesting" subgroup comparisons that help us gain more knowledge with multi-attribute groups?

Finally, though there exists interactive subgroup discovery algorithms (e.g., Dzyuba and van Leeuwen 2013) that help the user reduce uninteresting patterns on the fly of subgroup mining, the number of extracted patterns—in our case, subgroup comparisons—could still be large, preventing us from understanding the patterns on the meta-level. Our corresponding research question is:

> RQ5: How can we design an exploratory visualization tool that summarizes the comparisons and highlights the meta-patterns on extracted subgroup comparisons?

We examine the related work in the field of information visualization in relation to our design choices of the visualization tool in Sect. 6.

## 3 Data

### 3.1 Data collection and overview

During November 2013–January 2015, in a user study De Wolf et al. (2015), we collected 199 Facebook users' data with their consent. The data consist of friend graphs, user profiles, chats and posts. The box plot Tukey (1977) for the number of friends per user is shown in Fig. 1. In total, we identify 66,013 users with profiles, 49.2 % male, 50 % female and unspecified for the rest. 64.6 % of these users specify their birth dates, mostly people in their 20 s. Sixty-one percentage specified their home towns, out of whom 68 % come from Belgium, the rest are mainly from Spain (5 %), the Netherlands (3 %), Germany (3 %), Italy (2 %) and France (1 %). In both *chats* and *posts*, a user types a main

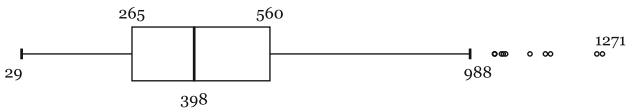


**Fig. 1** *Box plot* for the number of friends per user. The minimum, first quartile, median, third quartile and maximum are 29, 265, 398, 560 and 988, respectively. The 10 outliers are represented by *circles*

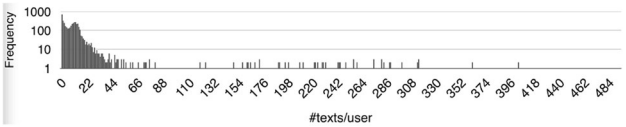


**Fig. 2** Histogram of #texts/user frequencies in *chats*

message or a comment to communicate, "text" for short. We use #texts/user to refer to the distribution of the number of texts sent by a user, #words/text and #chars/text to refer to the distributions of the number of words (space-separated) and the number of characters (UTF-8) in a text. Shapiro–Wilk tests Shapiro and Wilk (1965) on both the original values and log-scaled values of #texts/user, #words/text and #chars/text show that the respective distributions are significantly non-normal ($p < .001$). Indeed, we would expect exponential distributions here, such as Fig. 2. The median and IQR (interquartile range) values (IQR values in brackets) of #texts/user, #words/text and #chars/text are summarized in Table 1. We also see that the texts that people typed in *chats* and *posts* are short.

### 3.2 Language identification

In order to automatically detect the sentiments of the texts, we first need to sort the texts based on the languages in which they were written. However, language identification is non-trivial because of the corpus' large size, the many users from different countries and the short lengths of the texts. Marco Lui (2014) selected and compared eight language identification systems on labeled Twitter texts. They showed that an equal-weight voting over three systems consistently outperforms any individual system. These systems are: langid Lui and Baldwin (2012), LangDetect Nakatani (2011) and CLD2.[1] We adopt this method to identify the languages of the sentences in the corpus. HTML tags and URLs were removed beforehand. The results are summarized in Table 2. In total, 70,389 texts in 48 languages (83.1 % of the original texts) from *chats*, and 1,890,476 texts in 66 languages (86.6 % of the original texts) from *posts*, were identified. The languages of most texts in both *chats* and *posts* are Dutch, English, Spanish, German, French and Italian, as shown in Table 2. The

[1] https://code.google.com/p/cld2/.

**Table 1** Data summary of *chats* and *posts*

| User set | #Users | #Texts | #Texts/user | #Words/text | #Chars/text |
|---|---|---|---|---|---|
| *Chats* | 4480 | 84,751 | 10 (11) | 6 (9) | 23 (35) |
| *Posts* | 281,915 | 2,183,521 | 2 (3) | 6 (9) | 29 (41) |

**Table 2** Data summary for major languages

| Languages | *Chats* | | *Posts* | |
|---|---|---|---|---|
| | #Texts | #Users | #Texts | #Users |
| Dutch (nl) | 42,607 | 3268 | 400,349 | 73,497 |
| English (en) | 10,977 | 1894 | 635,997 | 117,521 |
| Spanish (es) | 1835 | 347 | 247,358 | 42,672 |
| German (de) | 4162 | 851 | 65,978 | 18,254 |
| French (fr) | 1086 | 425 | 38,952 | 10,745 |
| Italian (it) | 867 | 377 | 180,211 | 32,415 |

unidentified texts are usually very short phrases that are abbreviations, internet slang, (intentional) typos, emoticons and exclamation marks, such as "-_-||", ":)", "STUDY-YYYY!!!", "lmao!", or that occur in multiple languages such as "hehe," "amen". Eventually, we analyzed 74.1 % of the *chats* and 78.7 % of the *posts* in 11 languages.[2]

### 3.3 Attribute selection/construction

Each newsfeed post or chat record, with its comments, has an audience range, namely the set of (Facebook) users who can see the text. The texts in a chat are only visible to the chat participants. We can differentiate levels of privacy by the number of chat participants. The visibility of a text in a post is defined by its privacy setting, with four levels: *public*, *friends of friends* (*FoF*), *friends* and *custom*. The data statistics are summarized in Table 3. For profile attributes, we have "age" and "relationship status," as shown in Tables 4 and 5, respectively. We chose the age groups similarly to Stone et al. (2010). Moreover, since it is unlikely for people older than 80 years old to use Facebook, we assume these data to be untrustworthy and exclude the corresponding users from our analysis. Also, we find that 99.9 % of the users do not specify their "religion," "political-view" and "interested-in" features that are available in Facebook profiles.

### 3.4 Sentiment analysis

We use SentiStrength Thelwall et al. (2010a) to produce the texts' sentiment scores. It is a lexicon-based system that detects polarized sentiment strengths of short informal texts. It takes into account both terms and other language features such as booster words, negation, emoticons. Thelwall et al. (2010a) show that SentiStrength outperforms other common machine-learning algorithms. Abbasi et al. (2014) further show that the tool is generally better than other similar tools on five benchmark datasets. Because the term weights and language rules of SentiStrength are previously defined and no contextual texts are taken into account when predicting a text's sentiment, the positive and the negative scores of a text are generated independently and the positive/negative scores of different texts are generated independently. We run SentiStrength on the texts in *chats* and *posts*. Note that we do not conduct textual preprocessing beyond what SentiStrength provides. The counts of texts with positive and negative sentiment are summarized in Table 6. Notice that most chats and posts are neutral (value $\pm 1$), and negative sentiment occurs less often than positive.

## 4 Single-attribute sentiment differences: hypothesis testing

In this section, we test the sentiment differences according to *RQ1–RQ3*. Because the sentiment scores are highly skewed, we use the nonparametric Mann–Whitney test Mann and Whitney (1947) for two independent groups and Kruskal–Wallis test Kruskal and Wallis (1952) for $> 2$ independent groups. We report significant results[3] with two-tailed $p < .01$.

We exclude the texts with unspecified age or relationship status, and merge the [37, 50] and [51, 80] age groups in *chats* to account for larger group size. We test both positive ($s^+ \in [1, 5]$) and negative ($s^- \in [-5, -1]$) sentiment differences. Also, when needed, we test sentiment expressiveness ($s^+ - s^-$) differences. We will use the notation $\boldsymbol{GA} \, \mathbb{s} \, \boldsymbol{GB}$ with $\mathbb{s} \in \{>, <, \approx\}$ to denote that group $\boldsymbol{GA}$ is more, less than or similar to $\boldsymbol{GB}$ in terms of the absolute value of positive or negative sentiment, or expressiveness. Note that $\boldsymbol{GA} > \boldsymbol{GB}$ and $\boldsymbol{GB} \approx \boldsymbol{GC}$ do not automatically imply that $\boldsymbol{GA} > \boldsymbol{GC}$.

### 4.1 Sentiment differences between privacy levels (*RQ1*)

Tests show that the private chats and public posts differ significantly in positive sentiment ($U = 5.1 \times 10^{10}$), negative

[2] Namely English, Dutch, French, German, Italian, Polish, Portuguese, Russian, Spanish, Swedish and Turkish.

[3] Due to the limited space of this paper, we summarize and selectively report the results of the post hoc pairwise tests in Sect. 4. A complete report can be found at http://beaugogh.github.io/visualizations/mcells/data/pairwise.

**Table 3** Data summary for privacy levels

| #Participants | *Chats* | | Privacy setting | *Posts* | |
|---|---|---|---|---|---|
| | #Texts | #Users | | #Texts | #Users |
| 2 | 53,983 | 3249 | *Public* | 362,038 | 71,665 |
| [3, 4] | 4054 | 572 | *FoF* | 67,151 | 13,737 |
| [5, 6] | 1625 | 393 | *Friends* | 1,147,141 | 177,350 |
| [7, 10] | 1698 | 529 | *Custom* | 144,990 | 28,860 |
| [11, 20] | 1130 | 480 | | | |
| [21, 64] | 329 | 235 | | | |

**Table 4** Data summary for age

| Age | *Chats* | | *Posts* | |
|---|---|---|---|---|
| | #Texts | #Users | #Texts | #Users |
| [13, 16] | 544 | 59 | 1684 | 94 |
| [17, 20] | 15,516 | 754 | 65,676 | 1776 |
| [21, 24] | 15,627 | 816 | 236,128 | 4167 |
| [25, 28] | 7696 | 480 | 167,839 | 2454 |
| [29, 32] | 2744 | 208 | 79,763 | 984 |
| [33, 36] | 1405 | 99 | 36,088 | 399 |
| [37, 40] | 274 | 26 | 13,235 | 174 |
| [41, 50] | 419 | 54 | 24,629 | 303 |
| [51, 60] | 175 | 30 | 9484 | 144 |
| [61, 80] | 129 | 8 | 989 | 37 |

**Table 5** Data summary for relationship status

| Relation status | *Chats* | | *Posts* | |
|---|---|---|---|---|
| | #Texts | #Users | #Texts | #Users |
| Married | 626 | 36 | 81,029 | 856 |
| Relation | 4673 | 158 | 205,004 | 2462 |
| Single | 2818 | 190 | 196,107 | 2089 |

sentiment ($U = 5.2 \times 10^{10}$) and expressiveness ($U = 5.2 \times 10^{10}$). More specifically, the texts in *posts* are more positive and expressive than those in *chats*. The texts in *chats* are more negative than those in *posts*. This indicates that people tend to express more positive sentiment in *posts* shared with a broad audience, whereas they feel more free to express less positive, and also less extreme sentiments in *chats* that are exchanged within a private circle of participants. This partially confirms our hypothesis in *RQ1* that there is indeed a general pattern that people are more positive and less negative in public than in private on Facebook. Within *chats*, there is a difference between the groups of different privacy levels in positive sentiment ($\chi^2(5) = 29.0$), and negative sentiment ($\chi^2(5) = 83.5$). The conversations involving [11–20] participants are both more positive and negative than those involving 2 participants. It coincides with the general pattern that the texts are more sentimentally expressive in a more public setting. In *posts*, there is a difference in positive ($\chi^2(3) = 840.6$) and negative ($\chi^2(3) = 130.1$) sentiments between privacy levels: The *FoF* (friends of friends) texts are both more positive and negative than the texts with other settings. The texts with the *friends* and *custom* settings are more positive than the *public* texts. We can see that the texts with a "fairly public" setting (namely *FoF*) are more expressive than others, but the sentiments of the *public* texts are generally reserved.

### 4.2 Sentiment differences between ages (*RQ2*)

In *chats* and *posts*, there is a difference between age groups in positive sentiment $\left(\chi^2_{chats}(7) = 151.7,\ \chi^2_{posts}(9) = 4998.3\right)$, negative sentiment $\left(\chi^2_{chats}(7) = 123.1,\ \chi^2_{posts}(9) = 109.7\right)$ and expressiveness $\left(\chi^2_{chats}(7) = 99.6,\ \chi^2_{posts}(9) = 4403.0\right)$. Post hoc analysis reveals that younger people are generally more sentimentally expressive (Table 7). The [17, 20] group is also more negative than older age groups in *posts*, which supports the hypothesis in *RQ3* that negative sentiment

**Table 6** Summary of sentiment strength scores

| | Chats | | Posts | |
|---|---|---|---|---|
| | Positive | Negative | Positive | Negative |
| 1 | 43,305 (68.9 %) | 55,552 (88.4 %) | 1,074,092 (62.3 %) | 1,598,748 (92.7 %) |
| 2 | 18,317 (29.2 %) | 6664 (10.6 %) | 600,225 (34.8 %) | 102,543 (6.0 %) |
| 3 | 1117 (1.8 %) | 359 (0.57 %) | 44,910 (2.6 %) | 17,642 (1.0 %) |
| 4 | 96 (0.15 %) | 263 (0.42 %) | 5022 (0.29 %) | 5767 (0.33 %) |
| 5 | 7 (0.01 %) | 4 (0.006 %) | 572 (0.03 %) | 121 (0.007 %) |

declines after the early 20 s, but we do not see an increase in negative sentiment in the mid-age range [33, 59]. Interestingly, we see the opposite in *chats*: The [17, 20] group is less negative than [21, 50]. The late teen group seems to behave differently from older people in terms of negative sentiment expression. Younger people are generally more positive, which does not support the hypothesis that there is an increase in positivity after 50 years old.

### 4.3 Sentiment differences between relationships (*RQ3*)

Tests show that there is also a difference between groups with different relationship statuses, in positive sentiment $\left(\chi^2_{chats}(2) = 66.7,\ \chi^2_{posts}(2) = 2642.4\right)$ and negative sentiment $\left(\chi^2_{chats}(2) = 66.7,\ \chi^2_{posts}(2) = 303.0\right)$. More specifically, in both *chats* and *posts*, the texts from *single* users express more positive sentiment than those from *married* users. Also, the *posts* from *single* users express more negative sentiment than those from *married* users. It shows a contrast with *RQ3(a)* that *single* users actually express themselves more positively than *married* users. For *RQ3(b)*, in *chats*, *single* users express less negatively than users *in a relationship*. In *posts*, we find a stronger confirmation that users *in a relationship* express both less positive and more negative sentiment than *single* users. For *RQ3(c)*, users *in a relationship* have more positive chats and posts than those from *married* users. The users *in a relationship* also have more negative posts than the *married* users. These findings also show that *married* users are more neutral regarding online sentiment expression.

## 5 Multi-attribute sentiment differences: hypothesis exploration (*RQ4*)

So far, we have analyzed the sentiment differences between the groups of users defined by singular attribute values. It is straightforward to apply statistical tests in such scenarios. However, we often need to look into the "behavior" of user groups with combined attributes. For example, we find that the users with *married* relationship status tend to be less positive than the users with other statuses, but does this hold for both genders, different ages and so on? Existing approaches extract subgroups that have unusual or distinct distributional characteristics with respect to the entire population. For example, the target values in the subgroup "the 25- to 28-year-old males" are compared with the entire population. If this comparison produces a high score according to a certain quality measure, it is considered an interestingly distinct subgroup. However, instead of individual subgroups, we are interested in "comparisons" between subgroups, such as "the 25- to 28-year-old males" versus "the 29- to 32-year-old males," or "the 25- to 28-year-old males" versus "the males with the age interval other than 25–28."

**Table 7** Age group expressiveness in chats

| |
|---|
| [13, 16] > [21, 24] > [25, 28], [29, 32], [37, 50], [51, 80] |
| [17, 20] > [33, 36] > [21, 24] > [25, 28], [29, 32] |

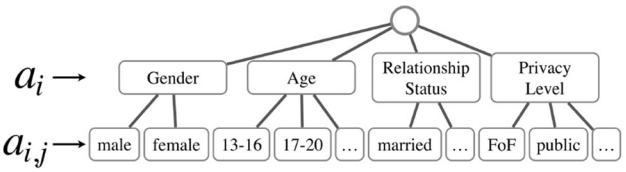


**Fig. 3** Illustration of hierarchy of attribute types and values

Furthermore, various quality measures are adopted or proposed to evaluate subgroups, and sometimes to prune the search space. But these measures often have a normality (Gaussian distribution) assumption for real-value target attributes (e.g., mean test, numeric weighted relative accuracy), whereas we see in Sect. 3, data could be non-normally distributed. We develop two top-down heuristic search algorithms, with statistical tests, without the normality assumption,[4] as both quality measures and pruning strategy, to extract subgroup comparisons. The algorithms are detailed in Sects. 5.2 and 5.3. These comparisons reveal interesting attribute combinations that provide a more fine-grained insight into the relationships between attributes and sentiments. They also offer potential sociological hypotheses for future study.

### 5.1 Notation

Consider a hierarchy $A$ of attribute types (labeled $a_i$) and values (labeled $a_{i,j}$), $i, j \in \mathbb{N}$, as shown in Fig. 3. Namely, $A = \{(a_i, A_i)\}, A_i = \{a_{i,j}\}$. We denote the complement of an attribute value $a_{i,k}$ within $A_i$ as $A'_{i,(k)} = \{a_{i,j} | j \neq k, a_{i,j} \in A_i\}$. Similarly, the complement of an attribute $a_i$ in the scope of $A$ is denoted as $A'_{(i)} = \{A_j | i \neq j, (a_j, A_j) \in A\}$.

Consider a subgroup $\boldsymbol{G}$ as a set of attribute values ($a_{i,j}$) where each corresponding attribute type ($a_i$) appears zero or one time. For example, a subgroup can be the males within 21–24 years old, namely {male, 21–24}. Let $\mathcal{G} = \{\boldsymbol{Gi}\}$ be the set of these subgroups. We use $\mathbb{m} \in M$ with $M = \{pos, neg, express\}$ to denote a chosen measure of positive sentiment, negative sentiment and expressiveness. We use the sign $\mathbb{s} \in \{>, <, \approx\}$, as defined in Sect. 4, to

[4] because of the usage of Mann–Whitney $U$ test.

describe the relationship between two subgroups, with the measure $\mathbb{m} \in M$, according to the statistical test $\mathbb{t}$ and the significance level $\alpha$. Let $\mathbb{t}(\boldsymbol{GA}, \boldsymbol{GB}, \mathbb{m})$ be the test that returns the sign $\mathbb{s}$ and the two-tailed $p$ value $p$, on subgroups $\boldsymbol{GA}$ and $\boldsymbol{GB}$ with the measure $\mathbb{m}$.

Note that the algorithms (Sects. 5.2 and 5.3) can be straightforwardly extended to accommodate a hierarchy with more levels of attribute values. For example, the age interval can be coarse initially, but divided into finer intervals at deeper levels.

### 5.2 Vertical comparisons

The algorithm for finding "vertical comparisons" of subgroups is detailed below. We use the set of comparisons $\mathcal{C}$ to store the comparisons between a target subgroup $\boldsymbol{G}$ (with $|\boldsymbol{G}| > 1$) and its "counterpart" $S$ (with $|\boldsymbol{G}| = |\boldsymbol{G} \cup S| + 1$), namely $\mathcal{C} = \{(\boldsymbol{G}, S, \mathbb{s}, \mathbb{m}, p)\}$. A depth-first search progressively accounts for subgroups with higher orders of attribute value combinations (lines 3, 4 and lines 14, 15). The significance level $\alpha$ serves as the pruning threshold that stops the search at a branch if the corresponding test's $p$ value is larger than $\alpha$ (lines 19–22).

```
1: Given A, m
2: C ← ∅, G ← ∅
3: for a_i do
4:     for a_{i,k} do
5:         G_i ← {a_{i,k}}
6:         s, p ← t(G_i, A'_{i,(k)}, m)
7:         C ← C ∪ {({a_{i,k}}, A_{i,(k)}, s, m, p)}
8:         CompareInDepth(G_i)
9:     end for
10: end for
11: function CompareInDepth(G_i)
12:     if G_i ∉ G then
13:         G ← G ∪ {G_i}
14:         for A_u ∈ A'_{(i)} do
15:             for a_{u,k} ∈ A_u do
16:                 G_u ← G_i ∪ {a_{u,k}}
17:                 S_u ← G_i ∪ A'_{u,(k)}
18:                 s, p ← t(G_u, S_u, m)
19:                 if p ≤ α then
20:                     C ← C ∪ {(G_u, S_u, s, m, p)}
21:                     CompareInDepth(G_u)
22:                 end if
23:             end for
24:         end for
25:     end if
26: end function
```

The algorithm finishes with a filled set of comparisons $\mathcal{C}$, which contains the comparisons of attribute combinations in different orders and their more general counterparts, informing us that by adding a specific attribute value, whether and how a combination is distinguishable from the rest. Table 8 shows the examples of vertical comparisons. For instance, we can see that the females of 21–24 years old express themselves less positively than other age groups in *chats*, but when such users chat in a group of 3–4 participants (i.e., privacy scope is 3–4), they express more positively than other age groups. Similarly, while the posts with the *friends* setting are generally less negative than those with other settings, the people of 25–28 years old express themselves more negatively in this setting.

### 5.3 Horizontal comparisons

While the vertical comparisons help us see the effect of adding/removing one attribute value on sentiment distributions, it is also desirable to see how different values of the same attribute affect sentiment distributions under different conditions. For example, how do {male, relation.},[5] {male, married}, {male, single} differ from each other? To this end, we present a second algorithm to extract horizontal comparisons, as shown below. Statistical tests are performed on a set of subgroups corresponding to all the attribute values $a_{u,k}$ under an attribute $a_u$ (line 17), conditioned on a previously given subgroup $\boldsymbol{Gi}$ (line 18). Each $a_{u,k}$ is added to $\boldsymbol{Gi}$ to form a more detailed subgroup, and all these subgroups are put through statistical testing, which returns significant comparisons (line 22). More formally, let $\mathcal{G}' = \{\boldsymbol{Gi}\}$ ($|\mathcal{G}'| \geq 2$) be a set of subgroups subject to post hoc analysis, and $\mathbb{t}(\mathcal{G}', \mathbb{m}, \alpha)$ the function that performs the pairwise testing and returns a set of comparisons that are significant at $\alpha$ level. Similar to the algorithm in Sect. 5.2, $\alpha$ serves as a threshold to remove the comparisons with large $p$ values.

Table 9 shows examples of horizontal comparisons. For example, from Sect. 4 we know that younger people are

[5] We use "relation." to denote the relationship status "in a relationship."

**Table 8** Examples of vertical comparisons ($\mathbb{m}$ is a sentiment measure, see Sect. 5.1)

| Data, $\mathbb{m}$ | *Comparison* | *Gender* | *Age* | *Relationship* | *Privacy* |
|---|---|---|---|---|---|
| Chats, pos | Age: 21–24 $<\neg$ 21–24 | Female | | | 3–4 |
| Chats, pos | Age: 21–24 $>\neg$ 21–24 | Female | | | |
| Posts, neg | Privacy: friends $>\neg$ friends | | 25–28 | | |
| Posts, neg | Privacy: friends $<\neg$ friends | | | | |

1: Given $A$, $\mathbb{m}$
2: $\mathcal{C} \leftarrow \emptyset$, $\mathcal{G} \leftarrow \emptyset$
3: **for** $a_i$ **do**
4: $\quad \mathcal{G}' \leftarrow \emptyset$
5: $\quad$ **for** $a_{i,k}$ **do**
6: $\quad\quad \boldsymbol{G}_i \leftarrow \{a_{i,k}\}$
7: $\quad\quad$ CompareInBreadth($\boldsymbol{G}_i$)
8: $\quad\quad \mathcal{G}' \leftarrow \mathcal{G}' \cup \{\boldsymbol{G}_i\}$
9: $\quad$ **end for**
10: $\quad \mathcal{C} \leftarrow \mathcal{C} \cup \mathbb{t}(\mathcal{G}', \mathbb{m}, \alpha)$
11: **end for**
12: **function** CompareInBreadth($\boldsymbol{G}_i$)
13: $\quad$ **if** $\boldsymbol{G}_i \notin \mathcal{G}$ **then**
14: $\quad\quad \mathcal{G} \leftarrow \mathcal{G} \cup \{\boldsymbol{G}_i\}$
15: $\quad\quad$ **for** $A_u \in A'_{(i)}$ **do**
16: $\quad\quad\quad \mathcal{G}' \leftarrow \emptyset$
17: $\quad\quad\quad$ **for** $a_{u,k} \in A_u$ **do**
18: $\quad\quad\quad\quad \boldsymbol{G}_u \leftarrow \boldsymbol{G}_i \cup \{a_{u,k}\}$
19: $\quad\quad\quad\quad$ CompareInBreadth($\boldsymbol{G}_u$)
20: $\quad\quad\quad\quad \mathcal{G}' \leftarrow \mathcal{G}' \cup \{\boldsymbol{G}_u\}$
21: $\quad\quad\quad$ **end for**
22: $\quad\quad\quad \mathcal{C} \leftarrow \mathcal{C} \cup \mathbb{t}(\mathcal{G}', \mathbb{m}, \alpha)$
23: $\quad\quad$ **end for**
24: $\quad$ **end if**
25: **end function**

more sentimentally expressive, as one base comparison {17–20}>{37–40} shows (second row in Table 9). However, when the privacy setting is *custom*, the expressiveness reverses, suggesting that the {17–20} group is not as expressive as they would be in a more public setting, and/or the {37–40} group expresses themselves more freely in a more private setting. Moreover, from Sect. 4 we know that in *posts*, the positive sentiment differences in relationship status are: {single}>{relation.}>{married}, but this pattern reverses when adding the "gender=male" attribute value, as shown in the table, providing us with a more differentiated view on the positive sentiment differences in relationship status.

### 5.4 Run-time performance

We take the datasets for chats and posts with available *gender*, *age*, *relationship* and *privacy* attribute values. The dataset for chats contains 6264 texts, and the dataset for posts contains 357,723 texts. We ran the algorithms for vertical and horizontal comparisons on both datasets, on positive sentiment scores. The settings of the computer are: Mac OS X, 2.9 Ghz Intel Core i5 processor, and 16 GB memory. Each algorithm was run 10 times on each dataset. The average running times and standard deviations (in seconds) are summarized in Table 10.

## 6 Exploratory visualization of subgroup comparisons (*RQ5*)

The number of subgroup comparisons produced by the algorithms in Sect. 5 can be very large. It thus becomes correspondingly difficult to examine the results. Information visualization is useful in helping data analysts quickly identify important patterns in large amounts of data, because the human visual system is highly parallel and preattentively sensitive to variations in visual stimuli, such as color, shape, positions. Shiffrin and Schneider (1977), Ware (2012).

Also, in Sect. 5 we mainly focus on a few subgroup comparisons showing contrasts. However, there are more ways in which we can examine the mined subgroup comparisons, and derive more insight. It is advisable to use various techniques to further filter and reduce the results so as to highlight the "essential bit." The outcome of this data analysis, especially when it is in an exploratory setting, depends on the data analyst to browse the results, test existing hypotheses and explore new patterns.

We developed an interactive visualization tool named *mCells* that is designed to visualize the mined subgroup comparisons, but is also applicable to a more general itemset visualization scenario. In this section, we first motivate our design choices in relation to existing works on analytic task taxonomy (Sect. 6.1) and itemset visualization (Sect. 6.2); we then detail the design of our tool named *mCells* (Sect. 6.3); lastly, we demonstrate the usefulness of the tool with case studies (Sect. 6.4).

**Table 9** Examples of horizontal comparisons ($\mathbb{m}$ is a sentiment measure, see Sect. 5.1)

| Data, $\mathbb{m}$ | *Comparison* | *Gender* | *Age* | *Relationship* | *Privacy* |
|---|---|---|---|---|---|
| Posts, express | Age: 17–20 < 37–40 | | | | Custom |
| Posts, express | Age: 17–20 > 37–40 | | | | |
| Posts, pos | Relationship: married > single | Male | | | |
| Posts, pos | Relationship: married < single | | | | |
| Posts, pos | Relationship: relation. > single | Male | | | |
| Posts, pos | Relationship: relation. < single | | | | |
| Posts, pos | Relationship: married > relation | Male | | | |
| Posts, pos | Relationship: married < relation | | | | |

### 6.1 Analytic tasks

Before we motivate our visualization design choices for itemset analysis, we need to understand the users' visual analytic tasks—what do users intend to achieve with information visualization tools in an analytic setting? Amar et al. (2005) proposed a taxonomy based on the data analysis tasks solicited from users. Yi et al. (2007) proposed a similar but more abstract interaction taxonomy based on the notion of user intent. Another user-oriented visual task taxonomy Zhou and Feiner (1998) targets a broad range of visual discourse that is not limited to information visualization. Some of its tasks, such as *identify* and *locate*, are suitable for scientific visualization in a non-analytic setting. Other taxonomies were geared more toward system-level interaction techniques Yi et al. (2007). Therefore, we consider the two taxonomies in Amar et al. (2005) and Yi et al. (2007) in the context of itemset visualization, as summarized below:[6]

**Retrieve value/ *select**: find and/or mark items with specific values.
**Filter**: find items that satisfy given conditions.
**Compute derived value**: compute aggregate numeric properties of a given itemset.
**Find extrema**: find the items with extreme attribute values.
**Determine range**: find the extreme values within a given itemset.
**Characterize distribution**: characterize the distribution of the values of a given itemset.
**Find anomalies**: identify anomalies with respect to certain metrics in a given itemset.
**Sort**: rank items according to certain metrics.
**Cluster**: find grouping(s) of a given itemset.
**Correlate/ *connect**: determine useful relationships between items or itemsets.
***Abstract/ *elaborate**: show more or less detail.

We make the following adjustments: (1) We put the tasks *Retrieve* and *Select* in the same category because both tasks involve the user "picking out" a specific data record. (2) As both taxonomies contain the task *Filter*, they are merged into one category. (3) We do not consider the tasks *Reconfigure* and *Encode* in Yi et al. (2007) as they are abstract tasks that overlap with more specific tasks such *Find Extrema*, *Sort*, *Cluster*. (4) The *Correlate* task captures the situation in which the user tries to find the existence or the degree of correlation between two sets of values. However, this notion is narrow. The *Connect* task includes a wider notion of connectedness or association discovery, between any items or itemsets. We thus put *Connect* in the same category with *Correlate* to make it more general. (5) The *Abstract/Elaborate* task in Yi et al. (2007) emphasizes the important parts of information visualization design—overview and details-on-demand Shneiderman (1996). Though other tasks such as *Retrieve Value* and *Find Anomalies* may entail *Abstract/ Elaborate*, the task itself is essential in exploratory data analysis from a user interaction point of view. We thus include it in the taxonomy.

**Table 10** Run times of the algorithms for vertical and horizontal comparisons, with in total 20 attribute values for the chats dataset and 24 attribute values for the posts dataset

| Data | #Texts | *Vertical* $(\mu, \sigma)$ | *Horizontal* $(\mu, \sigma)$ |
|---|---|---|---|
| *Chats* | 6264 | 0.88 (0.04) | 3.15 (0.13) |
| *Posts* | 357,723 | 62.58 (2.50) | 100.52 (2.27) |

### 6.2 Design choices

It is natural to consider itemsets in a traditional tabular form, as shown in Tables 8 and 9. Each row represents a comparison instance, and each column represents an attribute. Commercial applications such as Microsoft's Excel, Apple's Numbers and Google's online sheets are typical

[6] The tasks that are unique in Yi et al. (2007) are marked with $*$.

tools that help users manipulate and explore itemsets in a traditional tabular form. These tools can help the user retrieve/select items, filter items, compute aggregate numeric properties, find extrema according to a given itemset, determine the range of a given itemset and sort items. However, tasks such as *Characterize Distribution*, *Find Anomalies*, *Cluster*, *Correlate/Connect* and *Abstract/Elaborate* become difficult or impossible to perform when the table is large.[7] These tasks are particularly important in exploratory data analysis. Therefore, new forms of visualization are created to provide users with more powerful ways to examine itemsets.

We observe that there are two types of visualizations for itemsets in general: node-link diagrams and tables. The former emphasizes associations among items. It typically uses a node (circle, rectangle, etc.) to represent an item or itemset, and a link (straight line, curve, etc.) to represent an association between two nodes. The latter emphasizes distinct or interesting items or itemsets. It typically uses tabular cells, in rows and columns, to represent items or itemsets, and highlights the interesting ones by color coding or arranging the positions of the cells. The two types of visualizations could be mixed.

Typical examples in the "node-link diagram" approach include FIsViz Leung et al. (2008a), its variant WiFIsViz Leung et al. (2008b) and Circos Krzywinski et al. (2009). FIsViz and WiFIsViz first establish a grid and then use each column position of the grid to represent an item type, and each row position of the grid to represent the frequency of a certain item. Specific items, represented as nodes, are mapped onto the grid based on their column and row positions. Links are then drawn among the nodes to indicate associations. A series of nodes connected by *the same* link indicate an itemset. These tools can be helpful for the user to *characterize distributions*, *find anomalies*, *cluster* and *connect*. However, the drawbacks are: (1) links tend to occlude each other; (2) the links that connect different items can be difficult to recognize when different itemsets share the same items. Circos is a visualization tool that employs Chord diagrams[8] with the hierarchical-edge-bundling technique Holten (2006) to show relationships between entities. It is an effective tool that helps the user quickly identify the overall connection patterns in a diagram. But occlusion of links still occurs. Moreover, Chord diagrams condense all the data instances into one radial representation. When the number of data instances is large, it becomes difficult or impossible for the user to inspect an individual data instance. Figure 4 shows an example visualization generated with Circos. It visualizes 600 instances with five attributes C1–C5. There is a curve connecting each of the five attributes to an instance. We can see the small points mapped along the circumference and the occluded curves.

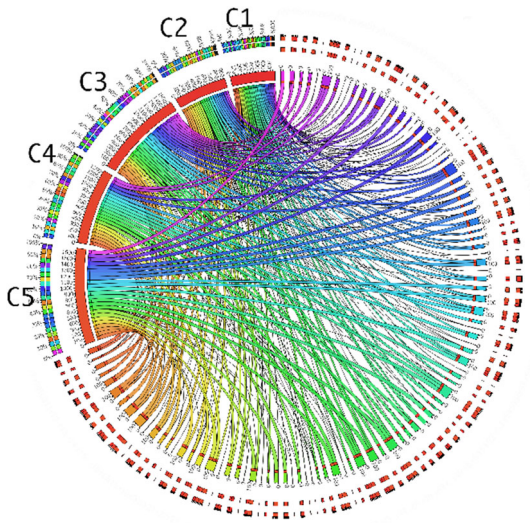


**Fig. 4** An example Chord diagram visualizing 600 rows with 5 attributes C1–C5, generated with Circos Krzywinski et al. (2009)

While the "node-link diagram" approach for visualizing itemsets is good for overview, the "tabular" approach is more suitable to ensure the visibility of individual data instances, because of its more efficient usage of space. A table can be spread over the entire screen. Compared with the "node-link diagram" approach, it is more easily scrollable and adjustable. Typical examples in the "tabular" approach include PowerSetViewer Munzner et al. (2005), TableLens Rao and Card (1995), Diversity Map Pham et al. (2010) and LineUp Gratzl et al. (2013). PowerSetViewer first indexes each incoming itemset with a horizontally one-dimensional array of cells, which, from left to right, increases the itemset order. For instance, the leftmost cell represents $\{a\}$ and the rightmost cell represents $\{a, b, c, d, e\}$. It then splits the one-dimensional array into lines of a grid, so that the top-left of the grid has relatively simple itemsets and the bottom-right of the grid contains more complex itemsets. The merit of this application is that it is part of a frequent itemset mining system where the user interactively queries the discovered itemsets. The tabular cells can be dynamically and efficiently constructed. However, this grid layout does not reflect the relationships between different items. It is thus difficult to inspect item distributions, clusters and associations.

TableLens, Diversity Map and LineUp all base their visualizations on the traditional tabular form, where a row represents a data instance, and a column represents an attribute. What these three tools have in common is that they introduce extra visual elements (colors and shapes) to make it easier for the user to spot distributions, discover

[7] The online article Krzywinski (2009) gives examples on the deficiencies of tables showing data.

[8] https://en.wikipedia.org/wiki/Chord_diagram.

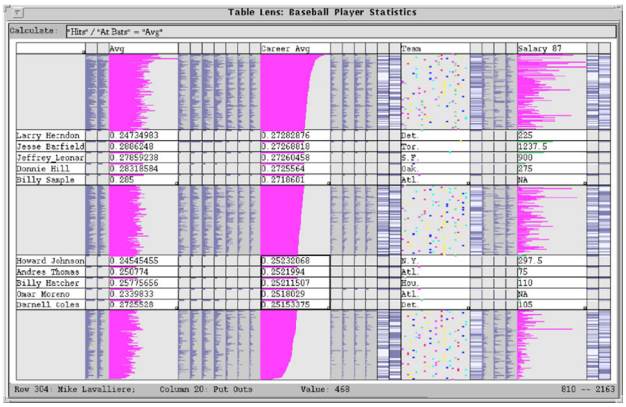


**Fig. 5** A screenshot of the original TableLens Rao and Card (1995)

clusters and make connections. TableLens introduces bars and colors to table cells. It enables comparison between the distributions of columns and reveals correlation. Diversity Map introduces gradient colors to encode the diversity of a table column. LineUp introduces stacked bar charts to table cells and enables manipulation of table rows and columns so that the user may compare the rankings of a row according to different criteria. Moreover, the fisheye view Furnas (1986) or other "details-on-demand" interactions can be easily applied to a tabular representation. Figure 5 shows a screenshot from the original TableLens. We can see that the individual items can be inspected without loosing their contexts. Next, we detail our visualization tool *mCells*,[9] which also bases its design on a tabular representation. Different from the previous tools that emphasize visualizing distributions of numerical values (TableLens and LineUp) or diversity of nominal values (Diversity Map), *mCells* emphasizes visualizing distributions and groupings of nominal values, enables the user to compare attributes, draw connections between items. Its features are tailored toward analyzing subgroup comparisons.

### 6.3 Visualization with *mCells*

The input format of *mCells* is similar to that of Tables 8 and 9. An artificial dataset is shown in Table 11, in which the leftmost column stores the comparisons, and the remaining columns store the attributes values on which the corresponding comparisons are conditioned. Figure 6 shows a screenshot of *mCells* visualizing a dataset similar to that in Table 11. The user can expand/shrink rows by dragging mouse downwards/upwards, so as to inspect individual rows without losing context (similar to TableLens). The user can also manually adjust the width/height of a column/row by dragging one of its edges.

[9] http://beaugogh.github.io/visualizations/mcells/.

**Table 11** Example dataset as input for *mCells*

| *Comparison* | *A* | *B* | *C* | *D* |
|---|---|---|---|---|
| B:B3 > NOT B3 | A1 | | C2 | D2 |
| B:B3 < NOT B3 | | | C1 | |
| A:A1 > NOT A1 | | | | D1 |
| A:A1 > NOT A1 | | B4 | C2 | D1 |
| A:A1 > NOT A1 | | B4 | | D1 |

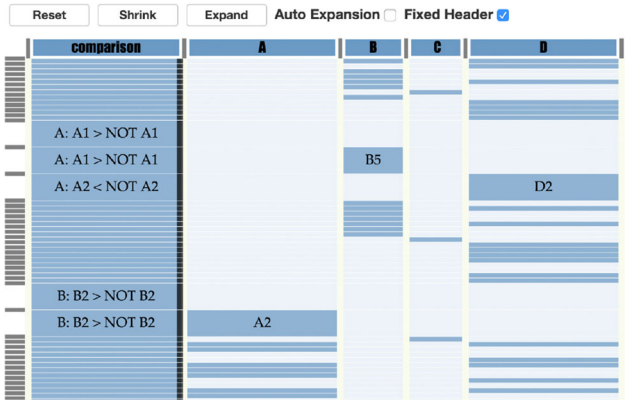


**Fig. 6** An overview of *mCells*, rows and columns may be expanded or folded

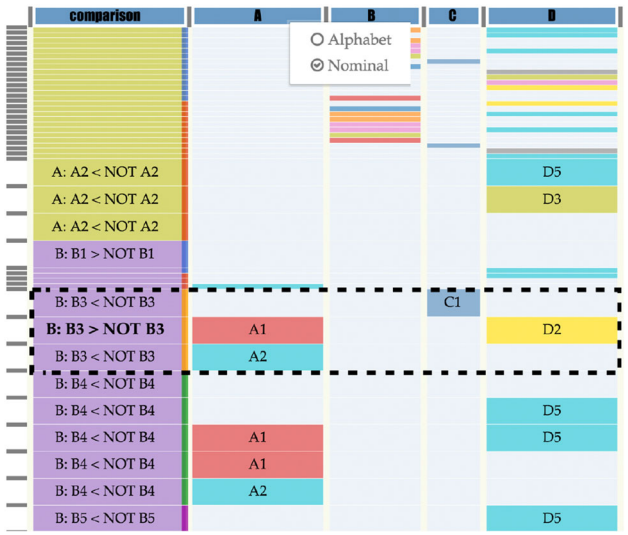


**Fig. 7** *Color* encoding of *mCells*, with *primary sort* on the comparison column and *secondary sort* on the A column

The user colors the column by selecting one encoding criterion from the header menu: alphabet (default) or nominal, as shown in Fig. 7. The *comparison* column is colored with respect to the attribute categories, and the vertical ribbon at the right of each cell in the *comparison* column is then colored with respect to specific types of comparisons. For example, as enclosed by the dashed lines, the three rows are about comparing the subgroup *B*3 with

*NOT B*3 on different conditions; thus, the small vertical ribbon to the right of each comparison cell is filled with the same orange color. Meanwhile, *mCells* detects the contrast comparison(s) within this sub-category and we see that $B : B3 > NOT\ B3$ is highlighted with bold font, as it is different from the other two comparisons.

Furthermore, the user may alphabetically sort the items in a column by clicking the column's header and we call this action *primary sort*. When necessary, the user may perform a *secondary sort* on a different column by holding down a key (e.g., ALT) and clicking the corresponding column's header. Figure 7 shows that the rows are primarily sorted according to the *comparison* column alphabetically, and secondarily sorted according to the *A* column. We can see that the items *A1* and *A2* are orderly stacked within each comparison category. For example, we can quickly see that, out of five the comparisons on *B*4 (purple cells with green ribbons), two involve *A1* and one involves *A2*.

*mCells* also provides an additional view for each type of comparison (e.g., $A : A1 > NOT\ A1$ in Table 11). The purpose is to enable the user to gain insight into the interplay between items of the same type of comparison locally. We first define the measure **s**, simply named "relative support," for each item and each link between two items (non-directional), as shown in Eq. 1. $S^{(\tau,i)}$ refers to the frequency of the item or link $i$ within a comparison type $\tau$, and $S^{(i)}$ refers to the total frequency of the item or link $i$.

$$\mathbf{s}^{(\tau,i)} = \frac{S^{(\tau,i)}}{S^{(i)}} \tag{1}$$

For example, as shown in Table 11, there are three instances for the comparison type $\tau = \{A : A1 > NOT\ A1\}$, within which $S^{(\tau,C2)} = 1$. The global frequency of *C*2 is $S^{(C2)} = 2$. Thus, the relative support of the item *C*2 is $\mathbf{s}^{(\tau,C2)} = 0.5$. Similarly, the relative support for the link $C2 - D1$ is $\mathbf{s}^{(\tau,C2-D1)} = 1$.

We position the items and links for each comparison type in a force-directed layout.[10] An item is directly represented with its text, and a link is represented with a straight line. We resize the font sizes of the items and the line weights of the links according to their relative support scores. Figure 8 demonstrates the layouts for two comparison types. When $A2 < NOT\ A2$, *D*4 appears frequently. Also, *B*4 and *D*5 often appear together. When $D5 < NOT\ D5$, *B*4, *A*1 and *B*3 often appear together.

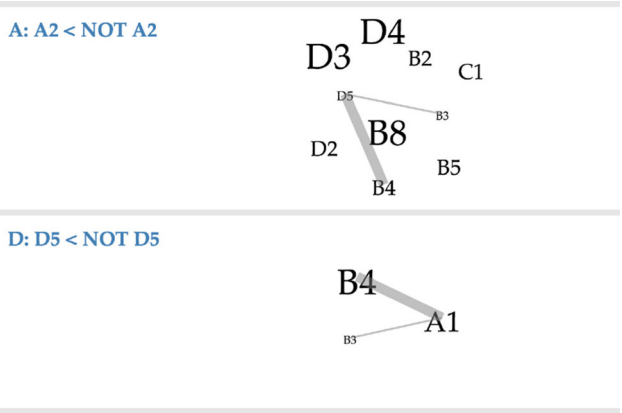


**Fig. 8** Force-directed layout for each comparison type, showing relationships among items

### 6.4 Case studies

In this section, we demonstrate the usefulness of *mCells* with the real datasets of extracted subgroup comparisons on user sentiment expression. We follow the analytic tasks summarized in Sect. 6.1. For the *retrieve value/select* task, as elaborated in Sect. 6.2, visualization tools based on a tabular form are suitable for retrieving or selecting specific individual items. We do not focus on the tasks *filter*, *compute derived value*, *find extrema* and *determine range*, because:

- Existing applications such as Excel and Numbers have widely adopted filtering functions, which includes searching with free texts and logical expressions;
- Computing derived values or determining extrema are for numerical values, whereas we focus on subgroup items with nominal values.

We then categorize the remaining analytic tasks into two categories: tasks as means, including *sort*, *cluster*, and *abstract/elaborate* and tasks as ends, including *characterize distribution*, *find anomalies* and *correlate/connect*. The "means" tasks are the intermediate steps to achieve the "end" tasks. We have introduced the interaction design of *mCells* in Sect. 6.3 that accommodates the tasks as means, namely

- *Sort*: alphabetical, primary/secondary sort;
- *Cluster*: nominal color encoding on columns and comparisons, and force-directed graph layouts[11] on different comparison types;
- *Abstract/elaborate*: foldable/expandable table rows and columns.

[10] https://github.com/mbostock/d3/wiki/Force-Layout.

[11] In a force-directed graph layout, heavily connected nodes form clusters.

Next, we instantiate the "end" tasks with concrete analytic questions on real datasets of extracted subgroup comparisons. More specifically, the questions are toward the vertical comparisons extracted on the users' positive sentiment expression in Facebook posts.

- *Characterize distribution*:

  Q1: How are the items distributed within the "relationship" attribute in all subgroup comparisons?
  Q2: How are the different comparisons distributed within the "privacy" attribute?
  Q3: How are the items of the "age" attribute distributed within the comparisons about *single* users versus non-*single* users?
  Q4: What are the main contextual items within the comparisons on the subgroups with the *friends* privacy setting?

- *Find anomalies*:

  Q5: Find contrast comparisons.
  Q6: What are the contexts of such comparisons, and how they are different from each other?

- *Correlate/connect*:[12]

  Q3 + Q4
  Q7: How do the items relate to one another within the same comparison type?

To address Q1, the user can first sort the "relationship" column alphabetically and fill the column cells with nominal colors, as detailed in the previous subsection. Figure 9 shows that the "single" relationship status participates the most in subgroup comparisons, followed by "married," "relation.", etc.

For Q2, the user may first sort and color the comparison column and then follow the "ribbon" colors (as explained in the previous subsection). We can see that many of the comparisons in the "privacy" category are about the *friends* privacy setting (ALL_FRIENDS), followed by the *public* setting (EVERYONE) and *custom* (CUSTOM), etc., as shown in Fig. 10.

For Q3, the user may first expand the corresponding comparison column cells, then "nominal-color" encode both the comparison column and the age column. Finally, the user performs a primary sort on the comparison column and a secondary sort on the age column. The result is shown in Fig. 11. We can see that the age group 17–20 appears the most often in the comparisons about *single* users, followed by 21–24, etc.

[12] The questions Q3 and Q4 also fall under this task description, because Q3 inquires about the relationship between two columns, and Q4 inquires about the relationship between a set of comparisons and their corresponding items.

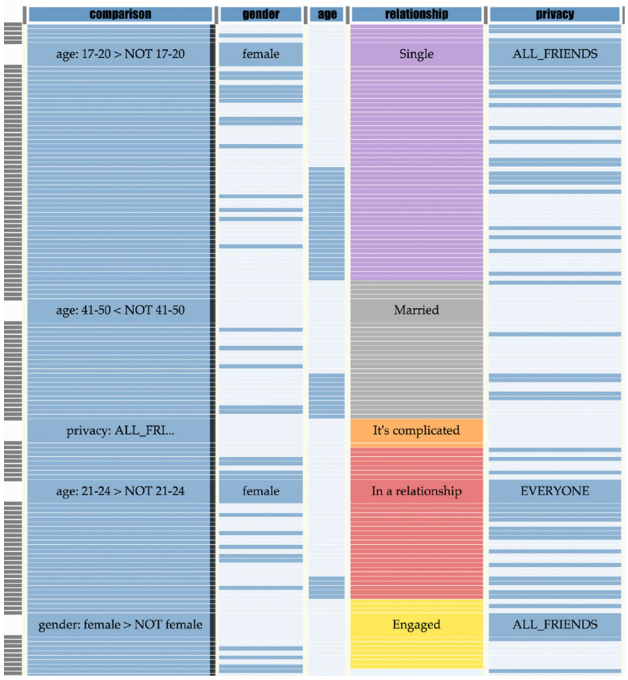


**Fig. 9** (Q1) The nominal *color* coding on the "relationship" column cells enables the user to see the distribution of different values


**Fig. 10** (Q2) The "ribbon" *colors* provide visual cues that show the distribution of different comparisons within the "privacy" attribute

For Q4, the user may switch to the graph view where he can inspect the main items associated with a comparison. From Fig. 12, we see that there exists a contrast pair of comparisons on the subgroups with the *friends* privacy setting. The "It's complicated" relationship status is the main factor appearing in the context of the first comparison (top). In other words, the posts with a *friends* setting tend to be less positive than those with other settings when a relationship is complicated. The "married," "Engaged" relationship statuses, the age intervals 41–50 and 13–16, etc. are the main factor appearing in the context of the second comparison (bottom). In other words, the posts with a *friends* setting tend to be more positive than those with


**Fig. 11** (Q3, Q5, Q6) A secondary sort shows the age item distribution within the comparisons about *single* users. Contrast comparisons are emphasized with *bold fonts*

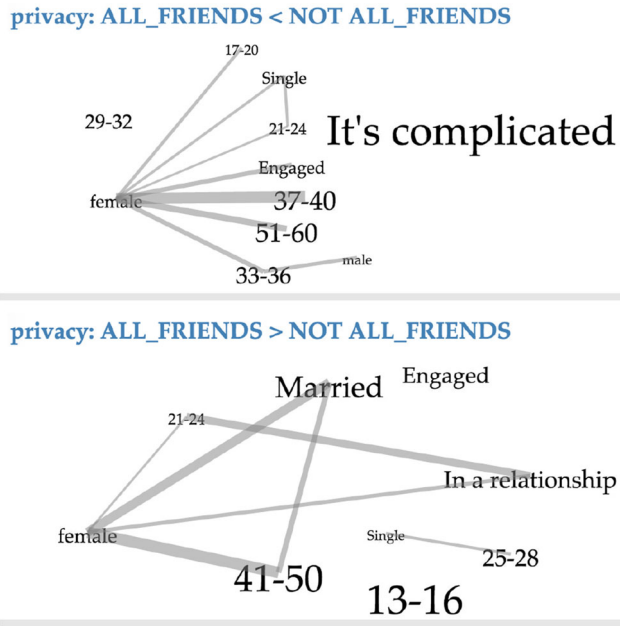


**Fig. 12** (Q4, Q6, Q7) Graph views show the resized contextual items and their connections of subgroup comparisons

other settings, when people are married, between 41 and 50, or every young, etc.

For Q5 and Q6, as shown in Fig. 11, contrast comparisons are emphasized with bold fonts. The user can directly compare how the corresponding contexts differ. For example, we can see that, out of the three emphasized comparisons (*single* < *NOT single*) in Fig. 11, two are distinctly with the age item 25–28. For the third comparison, the unique item combination of *female*, 17–20 and *FoF* contributes to its contrast. Similarly, we can also see a clear difference between the contextual items of the two comparisons in Fig. 12.

For Q7, we can see from Fig. 12 that, in both comparisons, the *female* item plays an important part for the comparisons to hold. The reason this item's font size is not enlarged is because the item also frequently appears elsewhere, making it less unique this the focused comparisons. We can also see that, in Fig. 12, in the first comparison (top), the *female* item is relatively strongly associated with the items 37–40, 51–60, etc. And in the second comparison (bottom), the *female* item is relatively strongly associated with 41–50. This shows that the *female* item with different age intervals contributes to different sentiment expression patterns.

More insight could be discovered via the visualizations of *mCells*.

## 7 Summary

In this paper, we take an interdisciplinary approach toward mining the patterns inherent to textual sentiments and metadata in online social networks.

We investigate the sentiment differences across privacy levels and demographic factors and find that not only the "conventional" or "stereotypical" hypotheses on demographic groups' sentiment expression are challenged, but also, importantly, that there are more detailed "stories" to be explored. For example, we find that the "coming of age" [17, 20] group wrote less negative texts in *chats* than older age groups, which counters our hypothesis that late teens have more negative texts.

Furthermore, while most social data analysis focuses on publicly available texts, we see different sentiment expressions from users under different privacy settings. It reminds us that people naturally adjust their communication with others according to the size of the audience, among many other factors. Investigating these differences will improve our understanding of the data. For example, we find that the texts posted publicly are in general more positive than those posted privately, but the texts with a complete *public* setting are more reserved.

Also, under the subgroup discovery paradigm, we present an approach with two algorithms that generalizes single-attribute testing, so as to provide more detailed insight into the relationships among different attributes, reveal interesting attribute value combinations with distinct sentiments and provide novel hypotheses for examination in future studies.

Finally, we design and develop an exploratory visualization tool named *mCells* that summarizes discovered subgroup comparisons and highlights meta-patterns, enabling the user to gain insight into subgroup comparison results.

## 8 Limitations and outlook

We apply statistical tests to identify differences between groups of sentiment scores, based on the assumption that each text's sentiment is independent of other texts'

sentiments. This assumption has two limitations: first, the sentiments of the texts from the same user may be correlated; second, the sentiments of the texts from the same chat or post may be correlated as well.

As seen in Sect. 3, the user sample in our dataset is biased. It consists of mostly young people from west European countries, particularly so for the users in *chats*, who are mostly Flemish students. Moreover, we only considered the users who have available profile features for demographic factors, which increases the bias.

Also, we used a tool (SentiStrength) to extract sentiment scores from the texts in multiple languages, which is bound to produce errors. Although it has been shown to be encouragingly accurate in relevant domains (Sects. 2, 3), it is yet to be investigated to which extent the inaccuracies may affect our results. We exclude the texts of which the language is unidentified. These texts include punctuations, emoticons and universal phrases, which account for a small proportion, but may still have an impact.

Furthermore, it is inherently difficult and ambiguous to rate a given sentence's sentiment. Often, people use negative words to be humorous or sarcastic, which could be counted as "positive." Sentiments also heavily depend on their contexts. Future studies can utilize context-based multi-dimensional sentiment analysis Scherer (2005).

Finally, it is worth investigating how *mCells* could help data analysts in real-life situations and how the tool could be improved accordingly.

**Acknowledgments** We thank Prof. Thelwall for his support with the SentiStrength tool. The research presented in this paper was supported by the Strategic Basic Research (SBO) Program of the Flemish Agency for Innovation via Science and Technology (IWT) through the project SPION (Grant No. 100048), and from the organization Fund for Scientific Research for Flanders (FWO) through the project Data Mining for Privacy in Social Networks (Grant No. G068611N).

## References

Abbasi A, Hassan A, Dhar M (2014) Benchmarking twitter sentiment analysis tools. In: The international conference on language resources and evaluation. pp 823–829

Amar R, Eagan J, Stasko J, (2005) Low-level components of analytic activity in information visualization. In: IEEE Symposium on information visualization, 2005. INFOVIS 2005. IEEE, pp 111–117

De Wolf R, Gao B, Berendt B, Pierson J (2015) The promise of audience transparency: exploring users perceptions and behaviors towards visualizations of networked audiences on facebook. Telemat Inform 32(4):890–908

Dzyuba V, van Leeuwen M (2013) Interactive discovery of interesting subgroup sets. In: Advances in intelligent data analysis XII. Springer, pp 150–161

Furnas GW (1986) Generalized fisheye views, vol 17. ACM, New York

Gratzl S, Lex A, Gehlenborg N, Pfister H, Streit M (2013) Lineup: visual analysis of multi-attribute rankings. IEEE Trans Vis Comput Graph 19(12):2277–2286

Gross JJ, Carstensen LL, Pasupathi M, Tsai J, Götestam Skorpen C, Hsu AY (1997) Emotion and aging: experience, expression, and control. Psychol Aging 12(4):590

Gross JJ, Richards JM, John OP (2006) Emotion regulation in everyday life. Emot Regul Couples Fam: Pathw Dysfunct Health 2006:13–35

Holten D (2006) Hierarchical edge bundles: visualization of adjacency relations in hierarchical data. IEEE Trans Vis Comput Graph 12(5):741–748

Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, pp 249–271

Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci 111(24):8788–8790

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621

Krzywinski M (2009) Circos visualizing tables, part I. http://circos.ca/presentations/articles/vis_tables1/

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645

Leman D, Feelders A, Knobbe A (2008) Exceptional model mining. In: Machine learning and knowledge discovery in databases. Springer, pp 1–16

Leung CKS, Irani PP, Carmichael CL (2008a) Fisviz: a frequent itemset visualizer. In: Advances in knowledge discovery and data mining. Springer, pp 644–652

Leung CKS, Irani PP, Carmichael CL (2008b) Wifisviz: effective visualization of frequent itemsets. In: Eighth IEEE international conference on data mining, 2008. ICDM'08. IEEE, pp 875–880

Lui M, Baldwin T (2012) langid.py: An off-the-shelf language identification tool. In: Proceedings of the ACL 2012 system demonstrations. Association for Computational Linguistics, pp 25–30

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60

Marco Lui TB (2014) Accurate language identification of Twitter messages. In: Proceedings of the 5th workshop on language analysis for social media (LASM)@EACL. pp 17–25

Munzner T, Kong Q, Ng RT, Lee J, Klawe J, Radulovic D, Leung CK (2005) Visual mining of power sets with large alphabets. Technical Report UBC CS TR-2005-25, Department of Computer Science, The University of British Columbia, Vancouver

Nakatani S (2011) Language detection library for Java. https://code.google.com/p/language-detection/

Pham T, Hess R, Ju C, Zhang E, Metoyer R (2010) Visualization of diversity in large multivariate data sets. IEEE Trans Vis Comput Graph 16(6):1053–1062

Rao R, Card SK (1995) Exploring large tables with the table lens. In: Conference companion on human factors in computing systems. ACM, pp 403–404

Scherer KR (2005) What are emotions? And how can they be measured? Soc Sci Inf 44(4):695–729

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611

Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychol Rev 84(2):127

Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. In: Visual languages, 1996. Proceedings., IEEE Symposium on. IEEE, pp 336–343

Siganos A, Vagenas-Nanos E, Verwijmeren P (2014) Facebook's daily sentiment and international stock markets. J Econ Behav Organ 107:730–743

Stone AA, Schwartz JE, Broderick JE, Deaton A (2010) A snapshot of the age distribution of psychological well-being in the United States. Proc Natl Acad Sci 107(22):9985–9990

Taylor J (2009) Emotional experience and romantic relationship status in emerging adult college women and men. Colorado State University, Fort Collins

Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010a) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61(12):2544–2558

Thelwall M, Wilkinson D, Uppal S (2010b) Data mining emotion in social network communication: gender differences in MySpace. J Am Soc Inf Sci Technol 61(1):190–199

Tukey JW (1977) Box-and-whisker plots. In: Exploratory data analysis, pp 39–43

van Leeuwen M, Knobbe A (2012) Diverse subgroup set discovery. Data Min Knowl Discov 25(2):208–242

Ware C (2012) Information visualization: perception for design. Elsevier, Amsterdam

Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Principles of data mining and knowledge discovery. Springer, pp 78–87

Yap SC, Anusic I, Lucas RE (2012) Does personality moderate reaction and adaptation to major life events? Evidence from the british household panel survey. J Res Personal 46(5):477–488

Yi JS, ah Kang Y, Stasko JT, Jacko JA (2007) Toward a deeper understanding of the role of interaction in information visualization. IEEE Trans Vis Comput Graph 13(6):1224–1231

Zhou MX, Feiner SK (1998) Visual task characterization for automated visual discourse synthesis. In: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co, pp 392–399